PR #25764 完整报告

sgl-project/sglang

[Codex] Remove stale DeepSeek V4 JIT kernels

合并时间：2026-05-19 20:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25764

执行摘要

本 PR 删除 DeepSeek V4 模型中不再使用的两个 JIT 内核（RMSNorm 和 SiLU-Mul-Quant），包括 Python 包装和 CUDA 实现，共减少 527 行代码。无功能变化，风险极低。

功能与动机

PR 作者发现 DeepSeek V4 的部分 JIT 内核已无任何代码引用（通过 rg 搜索验证），因此决定清理陈旧代码，减少维护负担和潜在混淆。引用 PR body：“remove the unused DeepSeek V4 rmsnorm JIT wrapper and kernel”以及“delete the stale silu_and_mul_masked_post_quant_tmp.cuh scratch kernel”。

实现拆解

1. 在 python/sglang/jit_kernel/deepseek_v4.py 中移除 @cache_once 装饰的 _jit_rmsnorm_head_module 函数及其调用者 rmsnorm_self 函数。这两个函数负责 JIT 编译并执行 RMSNorm 内核。
2. 删除 python/sglang/jit_kernel/csrc/deepseek_v4/rmsnorm.cuh 文件（134 行），包含 RMSNormKernel 结构体及其 run_self 方法。
3. 删除 python/sglang/jit_kernel/csrc/deepseek_v4/silu_and_mul_masked_post_quant_tmp.cuh 文件（371 行），包含复杂的分组 SiLU-Mul-Quant 内核函数。 所有删除集中在一个 commit 中完成，无任何测试或配置配套变更。

无，本 PR 仅删除代码，无新增或修改的内部逻辑。

评论区精华

无实质性讨论。Gemini Code Assist 自动评论无反馈，yuan-luo 直接批准。

风险与影响

风险极低。作者已通过 rg 确认无其他引用，且编译验证通过。影响范围仅限于清除无用代码，对功能无影响，对系统编译耗时略有减少。

关联脉络

本 PR 与 DeepSeek V4 系列内核优化相关，特别是 #26209 新增 FP4 Indexer 后，旧有的 RMSNorm 和 SiLU-Mul-Quant 临时内核不再被需要。清理工作为后续内核统一管理打下基础。